Stichprobengewichtung: ist Repräsentativität machbar?

Rothe, Günter; Wiedenbeck, Michael

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Stichprobengewichtung: Ist Repräsentativität machbar?

Bei der Analyse "repräsentativer" Stichprobenerhebungen werden zur Schätzung von Populationsmerkmalen in der Regel Gewichtungsvariablen herangezogen. Im folgenden Beitrag werden - sowohl theoretisch begründbare wie auch rein pragmatische - Ansätze zur Konstruktion von Gewichtsvariablen untersucht und die Probleme, die durch ihre Anwendung auftreten können, diskutiert: Die erwünschte Verbesserung erfolgt oft nur unter speziellen Modellannahmen, von deren Gültigkeit in der Regel nicht ausgegangen werden kann.

1. Vorbemerkung

Die Untersuchung großer Grundgesamtheiten mittels Zufallsstichproben ist inzwischen ein selbstverständlicher Standard für die empirischen Sozialwissenschaften geworden. Die Zulässigkeit von Verallgemeinerungen aus den Daten der Stichprobe auf die Grundgesamtheit hängt - so der allgemeine Sprachgebrauch - von der "Repräsentativität" der Stichprobe ab. Darunter wird verstanden, daß beliebige Merkmalsausprägungen in der Stichprobe im gleichen Anteil wie in der Grundgesamtheit, also "maßstabstreu", auftreten . Im Gegensatz zu diesem breiten sozialwissenschaftlichen Konsens kennt die mathematische Stichprobentheorie keine allgemein verbindliche Definition der "Repräsentativität", denn Stichproben sind grundsätzlich keine Substitute für die Grundgesamtheit. So ist z.B. die Zahl der denkbaren Antwortkombinationen bei nahezu jedem Fragenbogen in der Regel so groß, daß man selbst in der Grundgesamheit kaum zwei Befragte erwarten kann, die völlig identische Antworten geben würden. Damit kann natürlich erst recht nicht von einer Stichprobe erwartet werden, daß alle in der Gesamtpopulation auftretenden Antwortkombinationen "repräsentiert" sind.

Nichtsdestoweniger wird die "Qualität" einer Stichprobe oft daran gemessen, wie gut sie (ggf. unter Berücksichtigung von Gewichten) die Verteilungen spezieller soziodemographischer Variablen, die man aus anderen Erhebungen (z.B. Volkszählung oder Mikrozensus) genau zu kennen glaubt, widerspiegelt. Es wird dann erwartet, daß diese "Maßstabstreue" auch bei den anderen Variablen gilt, obwohl ihre Verteilungen nicht bekannt sind. Zumindest wird aber erwartet, daß die geeignete Verwendung von Gewichten gute Schätzungen von Populationsmerkmalen wie etwa Merkmalsdurchschnitten ermöglicht. Um die Berechtigung solcher Erwartungen zu überprüfen und die Rolle von Gewichten in Schätzverfahren vom mathematisch-statistischen Standpunkt her zu analysie-

ren, muß zunächst das Konstruktionsprinzip für die Gewichte genauer betrachtet werden und müssen die Eigenschaften der daraus resultierenden Schätzer untersucht werden.

Ein solches Schätzverfahren, das wir im nächsten Abschnitt behandeln werden, ist Grundvoraussetzung für die Verwendbarkeit der "Musterstichprobenpläne des Arbeitskreises Deutscher Marktforschungsinstitute". Dieses "ADM-Design" ist derzeit das in der Bundesrepublik für bundesweite Umfragen am meisten verwendete Stichprobenverfahren im Bereich der Marktforschung und der akademischen empirischen Sozialforschung.

Im dritten Abschnitt befassen wir uns mit einer Variante dieses Schätzers, bei der durch Modifikation der Gewichtsvariablen Maßstabstreue im obigen Sinn erzwungen wird. Diese kann als "nachträgliche Schichtung" interpretiert werden. Allen Nachgewichtungen, die von den dem ADM angeschlossenen Marktforschungsinstituten praktiziert werden, liegt dieses Vorgehen zugrunde. Die Praxis dieser Gewichtungen ist häufig Gegenstand der Kritik, insbesondere was die Transparenz des Verfahren hinsichtlich Genauigkeit und der Behandlung von Ausfällen anbelangt (vgl. etwa Kreienbrock 1983; Hanefeld 1982; Ullmer 1986).

Im vierten Abschnitt werden die Probleme diskutiert, die durch Antwortausfälle bei nachträglicher Schichtung auftreten.

Versuche, diese Schwierigkeiten zu lösen, bestehen nun in weiteren Modifikationen der Gewichtskonstruktion. Die Details der Lösungsverfahren sind in der Regel von Institut zu Institut verschieden und in der Regel nicht öffentlich zugänglich. Die Arbeitsgemeinschaft Media Analyse e.V. (AG.MA) verwendet jedoch ein Verfahren, das in seiner aktuellen Form in den MA-Dokumentationen 86 (S. 95ff) beschrieben ist. Es beruht auf einer iterierten Anpassung, deren Algorithmus detailliert bei Ockelmann (1981) dargestellt ist. Auch hier sind jedoch Ablauf und mathematischer Hintergrund nicht direkt zu ersehen, was für uns Anlaß war, dieses Konzept im fünften Abschnitt etwas sorgfältiger zu untersuchen, zu analysieren und einige seiner Eigenschaften herzuleiten.

2. Der HTQ-Schätzer

Das Standardvorgehen zur Schätzung eines Merkmalsdurchschnitts der Population aus dem Datensatz einer Stichprobe, der eine Gewichtungsvariable w und die Ausprägungen des Merkmals Y enthält, besteht in der Berechnung eines gewichteten Merkmalsdurchschnitts

$$\sum w_i Y_i \; / \; \sum w_i \; ,$$

wobei sich die Summation jeweils über alle Elemente der Stichprobe erstreckt. Um die Eigenschaften dieses Schätzers untersuchen zu können, ist zunächst etwas mathematische Terminologie erforderlich:

Wir gehen davon aus, daß eine Grundgesamtheit G vom Umfang N untersucht werden soll, der Einfachheit halber sei $G=\{1,\ldots,N\}$. Untersucht werden soll ein Merkmal Y, d.h. ein beliebiges Element i in G besitze das (reelle) Merkmal Y_i. Von Interesse ist in der Regel die Merkmalssumme $Y_\bullet = \sum_{1\leq i\leq N} Y_i$ oder der Merkmalsdurchschnitt $\bar{Y}_\bullet = Y_\bullet/N$, der aufgrund einer Stichprobe S, also einer Teilmenge von G, möglichst gut geschätzt werden soll. Mit n bezeichnen wir im folgenden den Stichprobenumfang $|S|$, also die Anzahl der Elemente in S. Ein Stichprobenplan besteht nun in einer Vorschrift, die den Zufallsmechanismus beschreibt, der schließlich die aktuelle zu erhebende Stichprobe zustandebringt. Damit kann ein Stichprobenplan als Wahrscheinlichkeitsverteilung P auf allen Teilmengen S von G aufgefaßt werden; P(S) gibt also die Wahrscheinlichkeit an, daß durch den Stichprobenplan gerade die Stichprobe S zustandekommt. Die Schätzung eines Merkmalswertes, etwa $Y_\bullet$, ist nun eine Zuordnungsvorschrift, die jeder denkbaren Stichprobe S einen Wert h(S) zuordnet und hierbei nur Merkmalswerte von Elementen aus S verwendet. Das Hauptqualitätskriterium ist hierbei in der Regel die Erwartungstreue, d.h. man verlangt

$$E(h)=\sum h(s)\cdot P(s)=Y_\bullet \; ,$$

wobei die Summation über alle Teilmengen s von G erfolgt. Im Prinzip ist dies das Kriterium, das die "Basisqualität" einer Stichprobe beschreibt.

Bezeichnen wir nun mit

$$T_i(S) = T_i = \begin{cases} 1 & \text{, falls } i \in S, \\ 0 & \text{sonst} \end{cases}$$

den Indikator für das Ereignis, daß Element i in die Stichprobe gelangt, so ist

$$p_i = P(T_i=1) = E(T_i)$$

die Wahrscheinlichkeit, daß i in die Stichprobe gelangt. Der klassische Horvitz-Thompson-Schätzer für die Merkmalssumme ist nun

$$\hat{Y}_o = \sum_{1 \leq i \leq N} Y_i \cdot T_i / p_i$$

(vgl. etwa Stenger 1986:201f). Man beachte, daß für die Berechnung nur die Merkmale und die Auswahlwahrscheinlichkeiten für diejenigen Elemente der Grundgesamtheit benötigt werden, die tatsächlich in die Stichprobe gelangen (in den anderen Fällen sind die T_i und damit die entsprechenden Summanden 0). Dieser Schätzer ist erwartungstreu für Y_o, zumindest solange für alle Einheiten der Grundgesamtheit $p_i>0$ gilt. Analog kann $\bar{Y}_o$ durch $\hat{Y}_o/N$ geschätzt werden, sofern der Umfang der Grundgesamtheit überhaupt bekannt ist. Ist dies nicht der Fall, so begnügt man sich häufig mit einer Schätzung

$$\hat{N} = \sum_{1 \leq i \leq N} T_i / p_i \ ,$$

die ebenfalls eine (erwartungstreue) Horvitz-Thompson-Schätzung ist (man betrachte den Spezialfall $Y_i=1$ für alle i!), und verwendet

$$\hat{\bar{Y}}_{(1)} = \frac{\sum_{1 \leq i \leq N} Y_i T_i / p_i}{\sum_{1 \leq i \leq N} T_i / p_i} \ .$$

Diesen Schätzer bezeichnen wir im folgenden als "HTQ(=Horvitz-Thompson-Quotienten)-Schätzer". Er hat allerdings den Nachteil, daß er als Quotientenschätzer nicht mehr erwartungstreu ist. Die Verzerrung liegt jedoch noch in vertretbarem Rahmen, solange die Schätzung $\hat{N}$ "hinreichend" wenig streut (wobei dies natürlich kaum quantifizierbar ist). Ein Vorteil dieses Schätzers besteht darin, daß nicht einmal die Wahrscheinlichkeiten p_i der in die Stichprobe gelangten Einheiten komplett, sondern nur bis auf eine Konstante

genau bekannt zu sein brauchen: Ist $p_i = q_i \cdot c(S)$ für alle $i \in S$, so gilt ebenfalls

$$\hat{\bar{Y}}_{(1)} = \frac{\sum_{1 \leq i \leq N} Y_i T_i / q_i}{\sum_{1 \leq i \leq N} T_i / q_i} .$$

Da das Gewicht $1/p_i$ nur bis auf einen konstanten Faktor wesentlich ist (die Multiplikation aller Gewichte mit einem konstanten Faktor ändert die Schätzung nicht), wird es in den meisten Fällen so zu $w_i = c/p_i$ normiert, daß

$$\sum_{i \in S} w_i = n ,$$

d.h. die Summe aller Gewichte in der Stichprobe genau den tatsächlich erreichten Stichprobenumfang entspricht.

Dieser Ansatz bildet die Grundlage des Designs von Zufallsstichproben wie etwa dem bereits angesprochenen ADM-Stichprobenplan. Es soll an dieser Stelle nicht mehr im Detail das Konzept der ADM-Designs beschrieben werden. Dies kann vom interessierten Leser etwa bei Kirschner (1980) nachgelesen werden. Von Bedeutung ist, daß das Verfahren schließlich zu einer im wesentlichen selbstgewichtenden Haushaltsstichprobe führen soll, d.h. man geht von der (aus verschiedenen Gründen nicht unproblematischen) Annahme aus, daß jeder Haushalt mit der gleichen Wahrscheinlichkeit in die Stichprobe gelangt. Wird nun eine Stichprobe nicht mit den Haushalten, sondern mit Personen als Erhebungseinheiten benötigt, so wird aus dem jeweiligen Haushalt dann nach einem einfachen Zufallsprinzip ("Schwedenschlüssel") eine Person ausgewählt und befragt. Dann führt das ADM-Design zu einer Personenstichprobe, bei der die Auswahlwahrscheinlichkeit einer Person umgekehrt proportional ist zur der (um die nicht zur Grundgesamtheit gehörenden Personen reduzierte) Größe des Haushaltes, in dem sie lebt. Diese Information genügt aber, um den HTQ-Schätzer zu berechnen; die reduzierte Haushaltsgröße kann somit als "Gewicht" w_i in diesem Schätzer fungieren:

$$\hat{\bar{Y}}_{(1)} = \frac{\sum_{1 \leq i \leq N} Y_i T_i \cdot w_i}{\sum_{1 \leq i \leq N} T_i \cdot w_i} .$$

Einen auf diesem Ansatz basierenden Gewichtungsfaktor werden wir im folgenden als 'theoretisches Gewicht' und seine Konstruktion in Anlehlung an die Terminologie der AG.MA "Transformation" bezeichnen (vgl. etwa AG.MA-Dokumentationen 1980:43).

3. Nachträgliche Schichtung

In Umfragen werden in der Regel verschiedene Variablen erfragt, deren Verteilung in der Gesamtpopulation zumindest theoretisch bereits bekannt ist: Eine Zuordnung zu Geschlecht, Alterklassen, Wohnregionen (also Regierungsbezirken etc.) erfolgt auch im Mikrozensus oder bei einer Volkszählung; sieht man die hieraus gewonnene Verteilung auf "Zellen" als "exakt" an, so lassen sich Abweichungen hiervon in der tatsächlich realisierten Stichprobe einerseits erklären durch die ohnehin zu erwartenden Zufallsfehler, andererseits aber auch durch Antwortausfälle, etwa dadurch, daß Personen zwar in die Stichprobe gelangten, aber nicht angetroffen wurden oder eine Antwort verweigerten. Es scheint nun naheliegend, unter- bzw. überrepräsentierte Zellen höher bzw. niedriger zu gewichten, um diesen Defekt auszugleichen. Ein solches Verfahren soll im folgenden detaillierter beschrieben werden. Wir gehen weiterhin davon aus, daß ein Merkmal Y erhoben wird, dessen Merkmalsdurchschnitt $\bar{Y}_o$ geschätzt werden soll. Darüberhinaus wird eine weitere Variable Z mit endlich vielen Ausprägungen $z_1,\ldots,z_K$ erhoben. Hier können natürlich auch mehrere Merkmale, wie etwa Geschlecht, Gehalts- und Altersgruppen kombiniert werden. Der Anteil S_k ("Soll") aller Elemente in der Grundgesamtheit mit der Ausprägung z_k, also

$$S_k = \text{Anzahl } (i: Z_i = z_k) \; / \; N \;,$$

von dem nun idealisierend angenommen werde, daß er durch externe Informationen vollständig bekannt sein möge, wird nun in Relation gesetzt zu dem entsprechenden Anteil I_k ("Ist") in der Stichprobe, also, wenn wir hierzu das im vorherigen Abschnitt eingeführte "theoretische" Gewicht mitberücksichtigen,

$$I_k = \Sigma_{[i:\, Z_i = z_k]} \, w_i T_i \;\; / \; \Sigma_{1 \le i \le N} \, w_i T_i \;.$$

Naheliegend ist nun eine Nachgewichtung nach dem Prinzip "Soll durch Ist" (SdI), also die Vergabe eines neuen Gewichts

$$\tilde{w}_i = w_i \cdot S_k / I_k \text{ , sofern } i \in G_k = \{i: Z_i = z_k\} .$$

Verwendet man nun in der HTQ-Schätzformel dieses neue Gewicht, also

$$\hat{\bar{Y}}_{(2)} = \frac{\sum_{1 \le i \le N} Y_i T_i \cdot \tilde{w}_i}{\sum_{1 \le i \le N} T_i \cdot \tilde{w}_i} ,$$

so zeigt eine einfache Rechnung

$$\hat{\bar{Y}}_{(2)} = \sum_{1 \le k \le K} S_k \cdot \hat{\bar{Y}}^{(k)}_{(2)} ,$$

wobei

$$\hat{\bar{Y}}^{(k)}_{(2)} = \frac{\sum_{i \in G_k} Y_i T_i \cdot \tilde{w}_i}{\sum_{i \in G_k} T_i \cdot \tilde{w}_i} .$$

$\hat{\bar{Y}}^{(k)}_{(2)}$ ist also genau der HTQ-Schätzer des Merkmalsdurchschnitts in der k-ten Zelle. Dieser Ansatz ist formal identisch mit den Schätzformeln bei geschichteten Stichproben, daher wird diese Art der Nachgewichtung oft als "nachträgliche Schichtung" umschrieben. In der Regel wird durch eine tatsächliche Schichtung eine Verbesserung der Schätzgenauigkeit erreicht; der Begriff "nachträgliche Schichtung" assoziiert daher, daß diese auch hiermit erzielt wird. Zwei Effekte sind hierbei allerdings zu berücksichtigen. Das größte Problem stellen dabei die Antwortausfälle dar, auf die wir später gesondert eingehen werden. Aber selbst in dem Idealfall, daß keine Ausfälle - welcher Art auch immer - zu verzeichnen wären, bleiben die einzelnen HTQ-Schätzer in den Zellen (wie bereits vorher der globale HTQ-Schätzer) verzerrt, und da die Stichprobenumfänge in den einzelnen Zellen kleiner sind, ist in der Regel auch eine stärkere Verzerrung zu erwarten. Dies bewirkt insbesondere, daß bei der Auswahl der Variablen die Aufteilung der Grundgesamtheit in Zellen nicht zu fein erfolgen kann, da in diesem Fall extrem kleine Zellenbesetzungen oder sogar Nullzellen in der Stichprobe realisiert werden. Es gibt Versuche, in solchen Fällen andere Gewichtungen vorzunehmen, die zwar auch auf dem SdI-Prinzip basieren, jedoch zusätzliche Dämpfungen

einbauen, die zu hohe Gewichte verhindern sollen. Statistische Rechtfertigungen für derartige Vorgehensweisen sind uns allerdings nicht bekannt.

4. Einflüsse von Antwortausfällen

Das im vorigen Abschnitt skizzierte SdI-Verfahren hat zum Ziel, die Verteilungen bestimmter erhobener demographischer Variablen anzupassen an bekannte Größen aus anderer Quelle. Werden diese Größen nicht zu fein strukturiert, sind die durch die speziellen Eigenschaften des HTQ-Schätzers auftretenden Verzerrungen tragbar, sofern keine Antwortausfälle auftreten würden. Diese jedoch stellen den Hauptgrund für die Durchführung einer Nachgewichtung dar. Paradoxerweise können dann jedoch die auftretenden Schätzfehler wesentlich gravierender sein. Um dies zu präzisieren, beschreiben wir das Antwortverhalten durch die externen Zufallsgrößen V_i derart, daß

$$V_i = \begin{cases} 1 & \text{, falls } T_i=1 \text{ und } Y_i \text{ erhebbar} \\ 0 & \text{sonst .} \end{cases}$$

Mit diesem Modellansatz lehnen wir uns an ein Konzept an, wie es etwa bei Oh und Scheuren (1983) verwendet wird. Für jedes $i \in G$ sei die individuelle, aber unbekannte Responsewahrscheinlichkeit

$$r_i = P(V_i=1 | T_i=1) ,$$

d.h. r_i ist die bedingte Wahrscheinlickeit, daß das Merkmal Y an der Einheit i erhoben werden kann, gegeben daß i in die Stichprobe gelangt ist. Damit gilt

$$E(V_i) = P(V_i=1 \text{ und } T_i=1) = P(V_i=1|T_i=1) \circ P(T_i=1) = r_i \circ p_i .$$

Die realisierte Stichprobe reduziert sich somit auf die Einheiten, für die $V_i=T_i=1$ gilt, beim exakten HTQ-Schätzer wäre T_i durch V_i und p_i durch $p_i \circ r_i$ zu ersetzen. Über r_i ist jedoch in der Regel nichts bekannt. Dies führt in der Praxis häufig dazu, so zu tun, als gäbe es keine Ausfälle, d.h. es wird lediglich T_i durch V_i ersetzt und man berechnet anstelle von $\hat{\bar{Y}}_{(1)}$ tatsächlich die Größe

$$\hat{\bar{Y}}^{*}_{(1)} = \frac{\sum_{1\leq i\leq N} Y_i V_i / p_i}{\sum_{1\leq i\leq N} V_i / p_i} ,$$

womit Zähler und Nenner des Schätzers nicht mehr die Erwartungswerte $Y_\bullet$ bzw. N, sondern $\sum_{i\in G} r_i Y_i$ bzw. $\sum_{i\in G} r_i$ aufweisen. Auch bei einer SdI-Nachgewichtung ergibt sich ein entsprechender "geschichteter", falscher HTQ-Schätzer:

$$\hat{\bar{Y}}^{*}_{(2)} = \sum_{1\leq k\leq K} S_k \circ \frac{\sum_{i\,\epsilon\, G_k} Y_i V_i / p_i}{\sum_{i\,\epsilon\, G_k} V_i / p_i} .$$

Konkret gibt es nur wenige Sonderfälle, in denen keine Änderungen auftreten: Ist etwa $r_i = c > 0$, also identisch für alle i (was natürlich illusorisch ist), so gilt $\hat{\bar{Y}}^{*}_{(1)} = \hat{\bar{Y}}_{(1)}$; kann man davon ausgehen, daß zumindest in den Zellen die Erreichbarkeitswahrscheinlichkeiten konstant sind (also $r_i = c_k > 0$ für $i \epsilon G_k$ für alle k), so ist zumindest $\hat{\bar{Y}}^{*}_{(2)} = \hat{\bar{Y}}_{(2)}$. Gilt dagegen in der Gesamtpopulation $\sum_i r_i Y_i = N^{-1} \sum_i r_i \cdot \sum_i Y_i$, was in etwa der Vorstellung einer "Unkorreliertheit von Merkmal und Erreichbarkeit" entspräche (dies ließe sich unter Verwendung des Konzept der Superpopulationen auch mathematisch präzisieren), so sind zumindest die Quotienten der Erwartungswerte von Zähler und Nenner bei $\hat{\bar{Y}}^{*}_{(1)}$ und $\hat{\bar{Y}}_{(1)}$ identisch - aber immer noch nicht die Erwartungswerte der Schätzer selbst (denn die Verzerrung ist in beiden Fällen unterschiedlich). Sind derartige Bedingungen nicht erfüllt, so kann eine nachträgliche Schichtung die Schätzung theoretisch sogar erheblich verschlechtern.

Ist man also auf Umfragen angewiesen, so kann man eigentlich nur hoffen, daß die gewählte Zellenstruktur zur Durchführung einer nachträglichen Schichtung so beschaffen ist, daß die Ausfallwahrscheinlichkeiten für jede Einheit innerhalb einer Zelle in etwa gleich ist; je homogener diese Wahrscheinlichkeiten, desto näher kommt die tatsächlich verwendete Schätzung dem angestrebten HTQ-Schätzer. Diese Vorstellung zeigt aber auch sofort das Dilemma, in dem sich die Umfrageforschung an dieser Stelle befindet: Es ist zwar vorstellbar, daß eine Verfeinerung der Zellenstruktur im obigen Sinne homogenere Zellen konstruiert, bei einer sehr feinen Struktur ist dagegen die Verwendung des HTQ-Schätzers gar nicht mehr erstrebenswert, da er dann hochgradig verzerrt sein wird aufgrund der am Ende von Abschnitt 3 angeführten Überlegungen. Da es aber kaum Alternativen zu HTQ-ähnlichen Schätzverfahren

gibt, wird man also offenbar mit einer Art "Unschärferelation" leben müssen: Je gröber die zur nachträglichen Schichtung herangezogene Zellaufteilung, desto schlechter ist die Anpassung an den HTQ-Schätzer; ist die Aufteilung feiner und damit die Anpassung besser, ist der HTQ-Schätzer selbst sehr schwach.

5. Redressement in der MA '86

Es gibt Versuche, dem oben angesprochenen Dilemma dadurch zu entgehen, daß man Verfahren verwendet, die nicht auf einer SdI-Gewichtung mit einer sehr feinen Zellaufteilung basieren, sondern die stattdessen mit mehreren, dafür aber gröberen Zellaufteilungen arbeiten und das SdI-Konzept in geeigneter (pragmatischer) Weise modifizieren. Bereits bei Deming und Stefan (1940) etwa werden zwei Variablen (deren gemeinsame Ausprägungen nicht bekannt waren oder nicht berücksichtigt wurden) simultan zur Nachgewichtung herangezogen; dieses inzwischen als "raking" bekannte Verfahren ist etwa bei Oh und Scheuren (1983) beschrieben. Das ZUMA-Gewicht der ALLBUS-Stichproben setzt dagegen spezielle Modellannahmen über das Zustandekommen von Ausfällen voraus (vgl. hierzu etwa Kirschner (1980) und Erbslöh und Wiedenbeck (1987)).

Ein Nachgewichtungsverfahren, das in der deutschen Marktforschung Verwendung findet und das - wenn auch mit etwas Mühe - aufgrund von Dokumentationen nachvollziehbar ist, ist das der Medienanalyse. Bei dieser regelmäßig stattfindenden Umfrage zur Erhebung von Informationen über die Verbreitung von Medien in der Bevölkerung wird erheblicher Aufwand getrieben, um eine hohe Ausschöpfung der Stichprobe zu erreichen, es wird dort auch ein Wert erreicht (nämlich über 85%), den Sozialwissenschaftler in ihren Untersuchungen aufgrund des wesentlich beschränkteren Kostenbudgets nicht erhoffen können. Aus diesem Grund sind bei der MA die im letzten Abschnitt angesprochenen Probleme der Antwortausfälle hier nicht so gravierend. Dennoch wird eine Nachgewichtung vorgenommen. Uns ist nicht bekannt, ob dieses Verfahren (oder geringfügige Modifikationen davon) auch bei anderen Umfragen zum Einsatz gelangt, dennoch soll es - quasi exemplarisch und weil es das einzige ist, zu dem wir Zugang gefunden haben - beschrieben und einige mathematische Eigenschaften angesprochen werden.

Zum Redressement der Personenstichprobe wurden 6 demographische Variablen herangezogen und gruppiert:

(A): Alter (in 7 Altersklassen)
(S): Geschlecht (2 Klassen)
(H): Haushaltsgröße (6 Klassen: 1,2,3,4,5 und mehr als 5 Personen)
(R): Regierungsbezirk (31 Bezirke)
(G): Gemeindegrößenklasse (7 Klassen nach Boustedt)
(T): Befragungstag (7 Wochentage)

Dabei ist anzumerken, daß die letzte Variable hierbei natürlich eine Sonderstellung einnimmt: Während die anderen für jede Person der Grundgesamtheit eine feste Größe darstellen, muß der Wochentag auch bei fester Person als zufällig angesehen werden. Überhaupt ist es fraglich, ob die Wochentagsverteilung bei der Mikrozensusbefragung so ohne weiteres vergleichbar ist mit derjenigen der MA-Befragung: Der Tag und Zeitpunkt der Befragung hängt sicherlich auch davon ab, wer im Einzelfall die Daten gewonnen hat. Man fragt sich, ob die Berücksichtigung dieser Variablen einen anderen Nutzen haben kann als zu demonstrieren, daß das Redressementverfahren schließlich die demographischen Variablen des Mikrozensus anpaßt.

Wird das ADM-Design dagegen als Haushaltsstichprobe verwendet, so werden als demographische Variablen nur die Größen Regierungsbezirk, Gemeindegröße und Haushaltsgröße verwendet.

Bei der Berücksichtigung aller Ausprägungskombinationen ergäben sich somit die Tafeln HxRxG mit 6x31x7= 1302 Zellen bzw. AxSxHxRxGxT mit 7x2x6x31x7x7 = 127596 Zellen im Fall der Haushalts- bzw. der Personenstichprobe. Gerade hierbei entstehen viele Zellen, die in einer Stichprobe unbesetzt bleiben. Dies führt bei Verwendung des SdI-Verfahrens insbesondere bei der Personenstichprobe zu den bereits in Abschnitt 3 angesprochenen Problemen der erhöhten Verzerrung.

Beim MA-Redressement werden nun zwei Prozeduren verwendet, um diesem Problem aus dem Wege zu gehen:

(I) Es werden nur Teiltafeln - genauer: zwei- bzw. dreidimensionale Randverteilungen - berücksichtigt und nach einem (noch zu beschreibenden) Verfahren simultan verrechnet.

Im konkreten Fall des Jahres 1986 wurden folgende zwei- bzw. dreidimensionale Randtafeln verwendet:

(1): AxSxR ≙ 7x2x31=434 Zellen
(2): AxSxT ≙ 7x2x7 = 98 Zellen
(3): TxR ≙ 7x31 =217 Zellen
(4): GxR ≙ 7x31 =217 Zellen
(5): HxR ≙ 6x31 =186 Zellen
(6): HxG ^ 6x7 = = 42 Zellen.

(II) Es werden vor dem Start des Iterationsverfahrens Zellen zusammengefaßt, die auch in diesen Randtafeln noch niedrig besetzt sind, um auf diese Weise "Schichten" mit hinreichend großem realisiertem Umfang in der Stichprobe zu gewinnen.

Die Restriktion auf Teiltafeln ignoriert natürlich Informationen über Wechselwirkungen dritter und höherer Ordnung; sie macht daher implizit eigentlich zusätzliche Modellannahmen über diese höheren Wechselwirkungen, diese Annahmen werden aber nicht erwähnt; die nachträgliche Schichtenbildung durch Zusammenfassung von Zellen ist durchaus Willkür ausgesetzt und aufgrund der uns vorliegenden Unterlagen auch nicht reproduzierbar.

Durch die Verwendung von L Randtafeln (L=1 bzw. L=6 beim Haushalts- bzw. Personenredressement) erfolgen L verschiedene Partitionen der Grundgesamtheit G in K_ℓ Zellen $G_{\ell k}$, $k=1,\ldots,K_\ell$, jeweils für $\ell=1,\ldots,L$; die Zellen mögen die jeweiligen Sollgewichte $S_{\ell k}$ besitzen. Die Redressementprozedur geht von den durch die "Transformation" vorgegebenen Individualgewichten $w_i^{(0)}$ aus und konstruiert iterativ neue Gewichte $w_i^{(m)}$, $m=1,2,3,\ldots$, solange, bis eine vorgegebene Stabilität erreicht ist (das genaue Abbruchkriterium haben wir nicht eruieren können). Zu jedem Iterationsschritt m gehören ebenfalls neue "Ist"-Zellenanteile $I_{\ell km}$ (analog zu denen in Abschnitt 3 konstruierten I_k). Wir betrachten nun drei Iterationsverfahren: Im m-ten Iterationsschritt werde das Gewicht $w_i^{(m-1)}$ der i-ten Person verändert zu

$$w_i^{(m)} = w_i^{(m-1)} \circ \Pi_{1\leq\ell\leq L} \tau_{i\ell(m-1)}^{(j)}, \quad j=1,2,3 ;$$

die Verfahren unterscheiden sich also nur in der Wahl des Faktors τ. Einfachster Ansatz wäre eine modifizierte SdI-Prozedur, nämlich

$$\tau_{i\ell m}^{(1)} := S_{\ell k}/I_{\ell km} \text{ für } i\in G_{\ell k} .$$

Im Fall einer Tafel wäre dies genau die SdI-Prozedur. Eine frühere Version des MA-Redressements besteht in der Verwendung einer "Dämpfung" der Form

$$\tau^{(2)}_{i\ell m} := (1-D_{i\ell m})\cdot 1 + D_{i\ell m}\cdot \tau^{(1)}_{i\ell m} \text{ für } i\in G_{\ell k} ,$$

wobei der Dämpfungsfaktor

$$D_{i\ell m}=(|S_{\ell k}-I_{\ell km}|/(\max_k|S_{\ell k}-I_{\ell km}|))^{1/2}$$

stets zwischen 0 und 1 liegt und daher der Faktor $\tau^{(2)}_{i\ell m}$ höchstens genausoweit von der 1 entfernt ist wie $\tau^{(1)}_{i\ell m}$. Das aktuelle Verfahren verwendet sogar

$$\tau^{(3)}_{i\ell m} = (\tau^{(2)}_{i\ell m})^{1/2} ;$$

dieser Faktor ist sogar stets noch näher an 1.

Um eine Vorstellung von der Wirkungsweise dieser Algorithmen zu gewinnen, betrachten wir zunächst einmal den Spezialfall einer Tafel, wie er ja konkret beim Haushaltsredressement vorliegt und bei dem offenbar auch die dritte der obigen Versionen derzeit verwendet wird. Der Einfachheit halber lassen wir daher im folgenden den Index ℓ weg.

Beim ersten Algorithmus ist der erste Schritt das SdI-Verfahren, danach bricht das Verfahren ab, da keine weiteren Änderungen mehr erfolgen.

Als nächstes stellen wir Überlegungen dahingehend an, welche Eigenschaften die beiden anderen Verfahren zumindest haben sollten, damit sie als sinnvoll anzusehen sind. Hier ist zunächst zu fordern, daß das Verfahren konvergiert. Hierzu müssen zwar nicht unbedingt die Folgen $w_i^{(m)}$ für $m \to \infty$ für jedes i einen Grenzwert besitzen, da die Konstruktion keine Normierung bewirkt ($w_\circ^{(m)} = \sum_i w_i^{(m)}$ ist nicht unbedingt identisch mit dem Stichprobenumfang n), aber zumindest sollte die Konvergenz der standardisierten Gewichte $\tilde{w}_i^{(m)} := w_i^{(m)} / w_\circ^{(m)}$, $m \to \infty$ erwartet werden (wobei die erste Konvergenz die zweite impliziert). Der Grenzwert $\tilde{w}_i^{(\infty)}$ schließlich, der ja dann als das "redressierte" Gewicht fungiert, kann eigentlich auch nur dann sinnvoll sein, wenn nie $\tilde{w}_i^{(\infty)} = 0$ gilt, da sonst Information über die Einheit i völlig ignoriert wird; zumindest sollte man für jedes $k\leq K$ die Forderung $\sum_{i\in G_k} \tilde{w}_i^{(\infty)} > 0$ stellen.

Es läßt sich nun zeigen (vgl. Anhang), daß diese Forderungen im Falle der Algorithmen (2) und (3) nur mit einem einzigen Grenzwert $\tilde{w}_i^{(\infty)}, 1 \leq i \leq N$, verträglich sind, nämlich wiederum dem SdI-Gewicht! Dies bedeutet, daß die MA-Verfahren des Redressements offenbar im Endeffekt nichts anderes bewirken als die SdI-Gewichtung, sofern wir uns in der Situation einer Tafel befinden - wie dies etwa konkret beim Haushaltsredressement der Fall ist. Führen die drei Verfahren, die wir in diesem Abschnitt untersucht haben, zu unterschiedlichen Ergebnissen, so kann das unserer Ansicht nach eigentlich nur am Abbruchkriterium für den Algorithmus liegen: Wenn Ockelmann (1981) 50 Iterationen nennt, so bedeutet dies, daß die diskutierte Konvergenz vom "Ist" zum "Soll" nur sehr langsam erfolgt und geringe Veränderungen der Zellgewichte von einem zum nächsten Iterationsschritt noch nicht unbedingt das Erreichen des Grenzwertes bedeuten. Offenbar wird nur durch einen vorzeitigen Abbruch die erstrebte Dämpfung erzielt.

Das Verhalten der Algorithmen im Falle mehrerer Tafeln ist wesentlich schwieriger zu handhaben; unsere Untersuchungen sind hier derzeit noch nicht abgeschlossen. An künstlichen Zahlenbeispielen läßt sich jedoch beobachten, daß etwa durch den (derzeit aktuellen) Algorithmus (3) nicht (wie etwa im Fall einer Tafel) unbedingt im Grenzwert die "Soll"-Verteilungen der jeweiligen Zellen erreicht werden können und daß auf der anderern Seite auch extreme Abweichungen zwischen "Soll" und "Ist" gelegentlich ignoriert werden. Es muß daher die Vermutung geäußert werden, daß die Anwendung dieses Verfahrens nur dann keinen Schaden anrichtet, wenn bereits vorher "Soll" und "Ist" nicht mehr allzusehr auseinanderklaffen; seine Auswirkung auf Stichproben mit geringerem Umfang und niedrigerer Ausschöpfung ist nicht zu überblicken.

6. Schlußbemerkungen

Wir haben in dieser Arbeit auf einige Probleme bei der Durchführung von Gewichtungen hinweisen wollen. Um Mißverständnisse auszuschließen, sei an dieser Stelle noch einmal betont, daß es nicht unsere Absicht war, ein spezielles Redressementverfahren zu verdammen; immerhin hat uns erst eine Diskussion mit Friedrich Wendt (quasi dem Urheber des ADM-Designs) über Sinn und Unsinn von Redressementverfahren im Herbst vergangenen Jahres und sein Hinweis auf die Dokumentation des MA-Verfahrens veranlaßt, diese Problematik aufzugreifen. Unklar ist insbesondere die Größenordnung der wiederholt angesprochenen Verzerrungen; auch ist zu fragen, ob die Verwendung von Schät-

zungen für die Ausfallwahrscheinlichkeiten r_i unter geeigneten Modellannahmen zu besseren Schätzern für $\bar{Y}_.$ führen kann. Aufgrund unserer bisherigen Überlegungen sind wir jedoch derzeit hinsichtlich der Leistungsfähigkeit von Nachgewichtungen außerordentlich skeptisch.

Dieser Beitrag wurde von Günter Rothe und Michael Wiedenbeck verfaßt, die bei ZUMA die Statistikabteilung vertreten.

Literatur

Arbeitsgemeinschaft Media-Analyse e.V. (Hrsg.), 1980/81/86: Media-Analyse 1980/81/86. Frankfurt a.M.

Deming, W.E./Stefan, F.F., 1940: On a least square adjustment of a sampled frequency table when the expected marginal totals are known. Ann. of Mathematical Statistics 11:427-444.

Erbslöh, B./Wiedenbeck, M., 1987: Methodenbericht zum ALLBUS 1986. ZUMA-Arbeitsbericht Nr. 1987/04.

Hanefeld, U., 1982: Die 78er ADM-Stichproben - Eine kritische Beschreibung der bevölkerungsrepräsentativen Zufallsstichprobe für die BRD. Arbeitspapier Nr. 74 des SfB 3, Frankfurt a.M./Mannheim.

Kirschner, H.-P., 1980: ALLBUS 1980: Stichprobenplan und Gewichtung. In: K.-U. Mayer/P. Schmidt (Hrsg.), Allgemeine Bevölkerungsumfrage der Sozialwissenschaften. ZUMA-Monographien, Band 5. Frankfurt.

Kreienbrock, L., 1983: Stichprobenpläne bei Marktforschungsinstituten. Diplomarbeit, Abteilung Statistik der Universität Dortmund.

Ockelmann, E., 1981: Das Redressement der Media-Analyse als analytische Korrektur der Felderergebnisse. S. 93-106 in: AG.MA (Hrsg.), MA 81-Dokumentationen.

Oh, H.L./Scheuren, F.J., 1983: Weighting adjustment for unit nonresponse. In: W.G. Madow/I. Olkin/D.B. Rubin: Incomplete Data in Panel Surveys Vol. II: Theory and Bibliographics.

Stenger, H., 1986: Stichproben. Heidelberg/Wien.

Ullmer, F., 1987: Wahlprognosen und Meinungsumfragen: Der Orakelspruch mit dem repräsentativen Querschnitt. Bild der Wissenschaften 1:89-100.

ANHANG: Konvergenzverhalten des Redressement-Algorithmus bei einer Tafel

Es mögen $\tilde{w}_i^{(\infty)}, 1 \leq i \leq N$ existieren, für die gelte:

(1) $\tilde{w}_i^{(m)} \dashrightarrow \tilde{w}_i^{(\infty)}$ für $m \dashrightarrow \infty$,

(2) $\sum_{i \in G_k} \tilde{w}_i^{(\infty)} > 0$ für $1 \leq k \leq K$,

(3) $\sum_{i \leq N} \tilde{w}_i^{(\infty)} = 1$.

Wir zeigen, daß dies nur im Fall $\tilde{w}_i^{(\infty)} = \tilde{w}_i^{(0)} \tau_{i0}^{(1)}$ möglich ist:
Wir betrachten die Menge aller möglichen Gewichte

$W=\{w\epsilon R^N: w_i \geq 0, \sum_{i\epsilon G_k} w_i > 0\}$ sowie $\tilde{W}=\{w\epsilon W: \sum_i w_i = 1\}$.

Dann sind bei fest vorgegebenem "Soll"-Vektor $S=(S_1,\ldots,S_Z)'\epsilon R^Z$ auf natürliche Weise Funktionen $h^{(j)}$ und $\tilde{h}^{(j)}$ auf W bzw. $\tilde{W}$ vorgegeben derart daß

$$w^{(m)} := (w_1^{(m)},\ldots,w_N^{(m)})' = h^{(j)}(w^{(m-1)}) \quad \text{sowie}$$

$$\tilde{w}^{(m)} = \tilde{h}^{(j)}(w^{(m-1)})$$

jeweils für die Algorithmen vom Typ j=2 bzw.j=3 gilt. Da offenbar $\tilde{h}^{(j)}$ in beiden Fällen auf $\tilde{W}$ stetig ist, folgt sofort

$$\tilde{h}^{(j)}(\tilde{w}^{(\infty)}) = \tilde{w}^{(\infty)} .$$

Dies impliziert die Existenz von Konstanten c_j, j=2,3 derart, daß

$$c_j h^{(j)}(w^{(\infty)}) = w^{(\infty)} \text{ bzw. } c_j \tau_k^{(j)}(w^{(\infty)}) = 1 \text{ für } 1 \leq k \leq K$$

in naheliegender Terminologie. Nun gilt aber offensichtlich

$$I_k^{(\infty)} := \sum_{i\epsilon G_k} w_i^{(\infty)} = S_k \text{ für alle } k ,$$

denn sonst gäbe es k_1 und k_2 mit $I_{k_1}^{(\infty)} < S_{k_1}$ und $I_{k_2}^{(\infty)} > S_{k_2}$, und hierfür wäre dann

$$1 < \tau_{k_1}^{(j)}(w^{(\infty)}) = 1/c_j = \tau_{k_2}^{(j)}(w^{(\infty)}) < 1 ,$$

was ein Widerspruch wäre. Damit ist zunächst für jede Zelle die Aussage

$$\sum_{i\epsilon G_k} \tilde{w}_i^{(\infty)} = \sum_{i\epsilon G_k} \tilde{w}_i^{(0)} \tau_{i0}^{(1)}$$

sichergestellt. Da sich nun aber schließlich das Verhältnis von Gewichten innerhalb einer vorgegebenen Zelle über die Iterationen hinweg nicht ändert, ist der Beweis erbracht.